DEVELOPMENT OF PUNJABI AUTOMATIC SPEECH RECOGNITION SYSTEM BASED ON TONAL ACOUSTIC FEATURES

RESEARCH PROPOSAL

SUBMITTED

IN PARTIAL FULFILLMENT TO THE REQUIREMENT

FOR

DOCTOR OF PHILOSOPHY

IN

FACULTY OF PHYSICAL SCIENCES
DISCIPLINE OF COMPUTER APPLICATIONS

By

JASPREET KAUR SANDHU

(Enrollment No. 17414FPE04)

Under the supervision of

Dr. AMITOJ SINGH
Assistant Professor,
Department of Computational Sciences,
Maharaja Ranjit Singh Punjab Technical University, BATHINDA

MAHARAJA RANJIT SINGH PUNJAB TECHNICAL UNIVERSITY

BATHINDA (INDIA)

(April, 2019)

Scanned by CamScanner

Title: DEVELOPMENT OF PUNJABI AUTOMATIC SPEECH RECOGNITION SYSTEM BASED ON TONAL ACOUSTIC FEATURES

1 Introduction

The most prominent and natural means of communications among human beings with machine is the speech. With the help of speech, the human beings can communicate easily, can explain as well as can investigate their ideas in various fields of life. The main purpose served by the speech is to convey the message. Legal sequence units help to convey the message and constraints must be agreed by the sequence in certain language. Vocal folds help in the production of speech and the articulators modify the air. The state of speech is used to label events in the speech. The three states of speech are stated below.

- **Silence (S):** There is no production of speech.
- **Unvoiced (U):** Vibration of vocal cord is absent and periodic or random type of resulted speech waveform is obtained.
- **Voiced (V):** Periodically vocal cords vibrate when flow of air is from the lungs.

Natural language processing is computer science branch whose main focus is on developing the systems that permit computer to communicate with humans using the language spoken every day. Natural language processing provides the system that is efficient and can be utilized by the users in their native language in a spoken, written or in both. These features are provided by the speech processing systems. Automatic Speech Recognition (ASR) system plays an important role in modern communication. The ASR system carries out the conversion of speech signal into readable text, actions or notations.

ASR for non tonal languages like English, French, Japanese, and Czech is well matured. A huge amount of work has been done for tonal languages like Chinese, Thai, Vietnamese and Mandarin but the tonal languages like Punjabi, Mizo, Bodo, Yoruba and Hausa are still in their nascent stage. A language is said to be tonal if the meaning of the word can be changed with the help of pitch of the word. Currently, the Indian researchers are showing great interest for building the reliable ASR system for tonal languages like Punjabi, Mizo, Bodo, Hausa and Yoruba.

1.1 Punjabi Language

Punjabi language is a tonal language that comes under the family of Indo-European language. Punjabi is spoken in India as well as in Pakistan and is the 10th most spoken language of the world and is the 22nd official language of India. It is official language of

Scanned by CamScanner

Punjab. Punjabi in India is written in the Gurumukhi script whereas in Pakistan it is written in Shahmukhi script.

a. Punjabi Language Phonological Structure

The study of speech as well as their function within sound system of a specific language is called phonology. The basic unit of language is phoneme. Punjabi phoneme has consonants, vowels, auxiliary signs and many dialects like Majhi, Pwadhi, Potwari, Dhani, Doabi, Hindko, Malwai etc. But, Majhi is the prestigious dialect.

b. Punjabi Tones

In language when pitch differentiate lexical or grammatical meaning then it is called tone. In Indian Punjabi high, low, mid/level tones are observed. The tonal characters in Punjabi are ਭ(bh),ਢ(dh),ਧ(dh),ਘ(gh),ਝ(jh). The tonal characters are considered as multifunction phonemes because the pronunciation of these tonal characters changes as per their position in the pronounced word. Over two successive syllables the contour tone is realized in Punjabi language. Each of the tone of Punjabi is implemented by shift of pitch. The tone in Punjabi is raised from consonantal mutation. Punjabi tone is seen on only one of the syllable and can take place with stress on it. The vocal cords vibration frequency assist in finding the category of the tone generated. If the position of the vowel (ਇ,ਉ,ਅ) is at initial syllable then tone is moved to the second syllable. On stressed vowels, tone is reflected and while restriction is experienced in vocal cords during the pronunciation of words having low tone.

Non- availability of effective Punjabi ASR as well as regional relevance has motivated us to work in discovered yet unexplored areas of Punjabi ASR system.

2 Literature Review

In the field of ASR, extensive research work has been done by various researchers as well as by research institutes. The literature review of some of the tonal languages is stated below.

2.1 Chinese

Yang *et al.* (2012) used the combination of Hidden Markov Model and Conditional Random Field (HMM-CRF) for automatically generation of Chinese abbreviations as well as expanded the vocabulary expression for the voice search with the help of abbreviation model. 88.3% accuracy was achieved for abbreviation modeling and for full name 79.2% accuracy was obtained. Li and Wu (2015) evaluated deep Long Short-Term Memory (LSTM) on Chinese large telephone speech recognition activity. The Character Error Rate (CER) of 34.65% was obtained by Long Short-Term Memory with Output Projection (LSTM-OP) layer.

29/8/2019

Scanned by CamScanner

Li *et al.* (2015a) built Chinese speech recognition system with Long Short-Term Memory Projected (LSTMP) network architecture and Connectionist Temporal Classification (CTC). The LSTMP-CTC was compared with DNN-HMM on the basis of CER. The CER obtained by LSTMP-CTC was 13.9% and the CER was 12.6% for DNN-HMM. Li *et al.* (2015b) evaluated the performance of various units of acoustic modeling with Deep Neural Network (DNN) for Chinese Large Vocabulary Continuous Speech Recognition (LVCSR) and compared it with Gaussian Mixture Model (GMM). The best CER observed by Context Dependent (CD) phone for GMM and DNN was 21.75% and 14.92%, respectively. Zhou *et al.* (2018a) compared syllable based model as well as CI (Context Independent) phoneme model. The CER achieved by CI-phoneme model and syllable model was 30.72% and 28.77%, respectively. Zhou *et al.* (2018b) have done the comparison of CI phoneme, syllables, words, sub-words and characters with help of transformer on Chinese ASR activities. The CER for CI phoneme was found to be 30.65%.

2.2 Mandarin

Amodei *et al.* (2016) recognized the English and Chinese speech with the help of end -to-end deep learning approach and look ahead convolution approaches. 9.39% and 6.22% accuracy was obtained by Recurrent Neural Network (RNN) for noisy and clean environment, respectively. Chan and Lane (2016) evaluated online speech recognition using attention based model on English as well as on Mandarin. The Word Error Rate (WER) obtained for English was 33.0% and Mandarin character model got a CER of 59.3%. Huang *et al.* (2017) proposed encoder classifier framework in which RNN was used for modeling the tones in Mandarin and the comparison of RNN-DNN with GMM-HMM indicated that RNN-DNN gave improved tone classification accuracy. Shan *et al.* (2018) discussed Mandarin speech recognition on the basis of voice search activity with the help of attention based encoder-decoder model. The CER and Sentence Error Rate (SER) of 3.58% and 7.43% was obtained, respectively, on MiTV voice search. Zou *et al.* (2018) for Mandarin speech recognition explored CTC model as well as attention based encoder decoder model. The CER achieved by CTC and attention based model on DidiCallcenter dataset was 7.45% and 5.68%, respectively.

2.3 Thai

Theera-Umpon *et al.* (2011) recognized Thai phonemes that included 21 initial consonants, 18 vowels and 9 final consonants using HMM and Mel Frequency Cepstrum Coefficient (MFCC) and Mel Frequency with Perceptual Linear Prediction (MFPLP). The tonal accents of

Scanned by CamScanner

Thai were recognized with the help of Neural Network (NN). The MFPLP gave better result as compared to MFCC. Kertkeidkachorn *et al.* (2012) discussed the spelling speech recognition corpus called CU-MFEC for Thai and English. Spelling task obtained accuracy of 79.37% and accuracy achieved by connected task was 54.92%. Srisuwan *et al.* (2013) developed Thai tone speech recognition system based on Electromyography (EMG). Top five features out of 68 features on the basis of RES index were selected and NN obtained 80% accuracy for Thai tones. Hu *et al.* (2014) explored Mandarin and Thai ASR system by including tone features into Convolutional Neural Network (CNN) and compared CNN with DNN. The WER for Thai was seen as 33.19% and 35.16% for CNN and DNN, respectively, for the VoiceTra corpus. Srijiranon and Eiamkanitchat (2015) presented a comparison of Neuro-Fuzzy algorithm with Decision Tree, Bayes, GMM and Support Vector Machine (SVM) on Thai words in continuous and easy control noise and loud and unpredictable noise. The Neuro-Fuzzy achieved higher accuracy in continuous and easy control noise as compared to Decision Tree, Bayes, GMM and SVM. But in loud and unpredictable noise, SVM achieved higher accuracy as compared to Decision Tree, Bayes, GMM and Neuro-Fuzzy.

2.4 Vietnamese

Nguyen and Tran (2012) conducted a study to find out the influences of tonal co-articulation of particles in Vietnamese. The study resulted that tonal co-articulation in Vietnamese always occur at transition between the two adjacent syllables. Huy *et al.* (2014) presented Multispace Distribution (MSD)-HMM for recognition of Vietnamese on the basis of Vietnamese tonal phonemes. The MSD-HMM gave 80.80% of Word Accuracy (WA) and HMM got WA of 78.31%.Nguyen *et al.* (2015) proposed algorithm that did the transformation of any Vietnamese word into pronunciation of tonal phoneme. The WER for VoiceTra test was 27.73% and for Basic Travel Expression Corpus (BTEC) test, it was 9.14%. Nguyen *et al.* (2016) used Deep Bottle Neck Features (DBNFs) in Vietnamese LVCSR system as well as used the tonal features. The DBNFs gave 10.76% of WER and WER achieved by HMM-GMM was 15.56%.

2.5 Mizo

Govind *et al.* (2012) studied the pitch slope role and duration in synthesizing Mizo tones and obtained the correctness score of 84%, 63% and 98% for recognizing real speech for falling tone, low tone and rising tone, respectively. Sarma *et al.* (2015) proposed automatically method to detect level tone or contour tone in Mizo. Classification was done on the basis of

29/8/2019

Scanned by CamScanner

pitch values as well as on the basis of height parameters obtained from pitch values fitted to tone. Dey *et al.* (2017) discussed the building of speech corpus and phone recognition system for Mizo. DNN and Subspace GMM (SGMM) were used for the development of Mizo phone recognition system. The Phone Error Rate (PER) for DNN was 23.2% and it was 10.3% for SGMM.

2.6 Punjabi

Dua *et al.* (2012) used HMM for implementing Punjabi speech to text system for connected words in classroom environment and open environment. The WER of 5.9% and 8.35% was observed for classroom and open environment, respectively. Lata and Arora (2012) done the study of Punjabi tones with relation to orthographic characters and concluded that starting and middle position give falling tones. Rising tone was indicated when tonal character is at last position. Lata and Arora (2013) have done the analysis of /h/ sound in Punjabi with PRAAT as well as MATLAB software on Malwai, a dialect of Punjabi. Kaur and Singh (2016a) used PNCC (Power Normalized Cepstral Coefficients) on connected words to build Punjabi ASR using HMM. WER in noise free and noisy environment obtained was 16.28% and 32.08%, respectively. Kaur and Singh (2016b) used MFCC, Perceptual Linear Prediction (PLP), PNCC to build ASR for Punjabi connected words using HMM.WER obtained by MFCC, PLP and PNCC was noted as 13.19%, 16.28% and 16.28%, respectively, for noise free environment. Kadyan *et al.* (2017a) proposed heterogeneous feature extraction that had been used on the speech signal for neutralizing the effect of the various conditions of non–ideal environment. Various feature extraction techniques were used with hybrid HMM classifier like HMM with Genetic Algorithm (GA) and HMM with Differential Evolution (DE). The Majhi obtained higher accuracy than Malwai, Powadhi and Doabi. The DE+HMM classifier gave the highest accuracy among the HMM, GA-HMM for both, speaker dependent and speaker independent, with MFCC. Kadyan *et al.* (2017b) evaluated the performance of ASR system for Punjabi using the combination of HMM with DE and with GA. The maximum WA in noisy environment was obtained by Malwai and it was 84.96%, 83.26% and 78.44% for DE-HMM, GA-HMM and HMM, respectively. Guglani and Mishra (2018) used the Kaldi ASR system for the recognition of Punjabi speech. MFCC gave less WER than PLP for Punjabi continuous speech. Kadyan *et al.* (2018) reduced acoustic mismatch between training and testing conditions with the help of DNN-HMM and handled the over fitting issue of training data with DNN-HMM. Arora *et al.* (2019) studied pitch, intensity and fundamental frequency and their effect on Punjabi dialects.

Scanned by CamScanner

Table 1: Some Important Works on Tonal Languages

Author	Year	Language	Publication	Dataset	Feature Extraction	Classifier	LM/ENV.	Result
Theera-Umpon *et al.*	2011	Thai	Expert Systems with Applications	Sentences:3045 Speakers:15 Duration: 88 hrs 8 min.	MFCC MFPLP MFPLP Pitch	HMM NN	Speaker-Dependent Speaker-Independent Speaker-Dependent Speaker-Independent	Classification rate (%) Vowels 67.71 67.86 67.82 Tonal accents 89.69 89.14
Dua *et al.*	2012	Punjabi	Proceedings of the 4th International Conference on Advances in Recent Technologies in Communication and Computing	Words:30-35	MFCC	HMM	ENV. Classroom Open	WER (%) 5.9 8.35
Govind *et al.*	2012	Mizo	Proceedings of the Speech Prosody	Speakers:1 Duration: 40 min	-	-	ENV. Real	Correctness (%) Falling tone:84.6 Low tone:63 Rising tone:98
Kertkeidkachorn *et al.*	2012	Thai	Proceedings of the International Conference on Speech Database and Assessments	Speakers:100 Words:16748 Duration: 58 hrs10 min	MFCC	HMM	-	Accuracy (%) Connectedalphabets:79.39 Continuous spelling:54.92

Scanned by CamScanner

Yang *et al.*	2012	Chinese	Computer Speech and Language	Speakers:20 Abbreviations:2200 Organizationfullname :1945 Duration:150 hrs	MFCC	HMM-CRF	ENV. Clean	Accuracy (%) Abbreviation modeling: 88.3 Fullname :79.2
Hu *et al.*	2014	Thai	Proceedings of the Signal and Information Processing Association Annual Summit and Conference (APSIPA)	Mandarin HUB4: Read speech Duration: 27 hrs. Thai VoiceTRA: Network speech Duration:100 hrs	MFCC	CNN DNN CNN DNN	-	CER(%) 10.92 12.57 WER(%) 33.19 35.16
Huy *et al.*	2014	Vietnamese	Journal of Computer Science and Cybernetics	VOV Speakers:30 Words: 23424, Distinct syllables with tone: 4923 Duration: 19 hrs	MFCC MFCC+NCC+ AMDF_MSD	HMM MSD-HMM	-	WA(%) 78.31 80.80
Li and Wu	2015	Chinese	Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)	HKUST Corpus Mandarin Chinese telephonic conversation Duration: 150 hrs	Mel-Scale log filter bank coefficients	DNN LSTM-OP	-	CER(%) 38.42 34.65
Li *et al.*	2015b	Chinese	Neurocomputing	Hub-4NE corpus Broadcast news of Chinese Duration:30 hrs	MFCC	GMM DNN	CD Model Phones	CER(%) 21.75 14.92
Nguyen *et al.*	2015	Vietnamese	Proceedings of the International Conference Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation (OCOCOSDA/CASLRE)	VoiceTra test Words:803 Duration:36 min BTEC test Words:510 Duration:19 min	BNF	GMM-HMM	-	WER (%) 27.73 (VoiceTra test) 9.14 (BTEC test)

Scanned by CamScanner

Sarma *et al.*	2015	Mizo	Proceedings of the 16th Annual Conference of the International Speech Communication Association	Speakers :5 Syllable:4384 which contained 672 high tones 2254 low tones 808 falling tones 650 rising tones	-	-	ENV Real	DR(%) Without CE with CE 50.14 35.26 81.28 56.78 75.23 78.92 52.47 74.62
Srijiranon and Eiamkanitchat	2015	Thai	Proceedings of the 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)	Speakers:7 males and 7 females Words:9	PLP	Decision Tree Bayes GMM SVM Neuro-Fuzzy Decision Tree Bayes GMM SVM Neuro-Fuzzy	Continuous & Easy control Noise Loud & Unpredictable noise	Accuracy (%) 53.70 56.13 73.76 76.57 84.85 36.57 30.06 52.66 60.50 59.81
Amodei *et al.*	2016	Mandarin	Proceedings of the 33rd International Conference on Machine Learning (ICML)	Mandarin Labeled speech Words:11 M	End -to -end RNN	RNN	Noisy Clean	Accuracy (%) 9.39 6.22
Chan and Lane	2016	Mandarin	Proceedings of the Interspeech	For English Wall Street Journal For Mandarin GALE 104 hrs speech	End to end RNN	RNN	English Online CTC Online CTC+Pinyin	WER (%) 33.0 CER (%) 59.3
Chen *et al.*	2016	Mandarin	Speech Communication	iCALL corpus- Speakers:305 Words: 90841 Duration: 142 hrs English speakers French Speakers	-	-	-	Lexical tone accuracy (%) Tones 1 2 3 4 73.0 71.4 62.4 72.2 71.2 57 53.5 57.8

Scanned by CamScanner

Kaur and Singh	2016a	Punjabi	Proceedings of the 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)	ENV: Clean Speakers:3 Words:43 ENV :Noisy Speakers:3 Words:53	PNCC	HMM	ENV: Noise free ENV: Noisy	WER (%) 16.28 32.08
Kaur and Singh	2016b	Punjabi	Proceedings of the 2nd International Conference on Advances in Computing, Communications and Informatics (ICACCI)	Total Duration: 3hrs 15 min. ENV: Clean Speakers:16 Words:43 ENV :Noisy Speakers:12 Words:53	MFCC PLP PNCC MFCC PLP PNCC	HMM	ENV: Noise free ENV: Noisy	WER (%) 13.19 16.28 16.28 32.08 28.08 32.08
Nguyen *et al.*	2016	Vietnamese	Procedia Computer Science	Speakers:30 Words : 23424 Duration:19 hrs	MFCC	HMM-GMM DBNFs	--	WER(%) 15.56 10.76
Dey *et al.*	2017	Mizo	Proceedings of the India Council International Conference (INDICON)	For training Speakers:62 Words: 7394 Duration:5 hrs5 min For testing Speakers:19 Words: 1568 Duration: 1 hr 5min	MFCC	DNN SGMM	-	PER (%) 23.2 10.3
Huang *et al.*	2017	Mandarin	arXivpreprint arXiv:1711.01946	Speakers:160 Words:92243 Duration:110 hrs	LDA-MLLT	 RNN DNN-RNN GMM-HMM	 - MLE MMI	Tone classification accuracy (%) 70.7 70.4 75.7

Scanned by CamScanner

Kadyan *et al.*	2017a	Punjabi	International Journal of Speech Technology	Speakers:25 Words: 5000 Duration:32 hrs	MFCC	HMM GA-HMM DE-HMM HMM GA-HMM DE+HMM	Speaker Dependent Speaker Independent	Accuracy (%) 60.97 70.42 74.59 58.48 67.36 72.47
Kadyan *et al.*	2017b	Punjabi	IETE Journal of Research	Training Speakers:24 Words: 58700 Duration: 46hrs 40 min Testing Speakers:13 Words: 6100 Duration:5hrs30 min	MFCC	DE-HMM GA-HMM HMM DE-HMM GA-HMM HMM	Env:Noisy Env:Real	WA (%) 84.96(Malwai) 83.26(Malwai) 78.44(Malwai) 82.26(Malwai) 78.47(Majhi) 80.49(Malwai)
Guglani and Mishra	2018	Punjabi	International Journal of Speech Technology	Speakers:24 Numbers: 0-9 digits in Malwai dialect Duration: 240 hrs	MFCC PLP MFCC PLP	GMM-HMM	Monophone 3-gram Triphone 3-gram	WER(%) 31.0 33.1 21.2 22.7
Kadyan *et al.*	2018	Punjabi	International Journal of Speech Technology	Dataset1:Connected speech corpus Speakers:13 Connected words:21764 Dataset2:Continuous speech corpus Sentences:3611 Speakers:13	MFCC MFCC GFCC	DNN-HMM	-	WER(%) Dataset1 39.73 Dataset2 5.22 24.67
Shan *et al.*	2018	Mandarin	Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)	Words: 4M approx. Duration:3000 hrs	Mel-scale filter bank	BLSTM	-	CER(%) SER(%) 3.58 7.43

Jaspreet Kaur
29/8/2019

Scanned by CamScanner

Zou et al.	*2018*	*Mandarin*	arXiv preprint arXiv:1805.03832	DidiCallcenter Spontaneous data Words:2.2 M Duration: About 2800 hrs DidiReading Reading style data Words:16.2M Duration: About 12000 hrs	Mel-scale filter bank	BLSTM LSTM	Attention model CTC Attention model CTC	CER(%) 5.68 7.45 4.89 5.79
Zhou *et al.*	2018a	Chinese	arXiv preprint arXiv:1804.10752	HKUST corpus Duration:150 hrs Chinese telephone speech	Log-Mel filter bank	-	CI-phoneme-D1024-H16 (Speed perturb) model Syllable-D1024-H16 (Speed perturb)	CER(%) 30.65 28.77
Zhou *et al.*	2018b	Chinese	Proceedings of the International Conference on Neural Information Processing (ICONIP)	HKUST corpus Chinese telephone speech Duration:150 hrs	Log-Mel filter bank	-	D1024-H16 (Speed perturb) Model CI phoneme Syllables Words Sub words Characters	CER(%) 30.65 28.77 27.42 27.26 26.64
Arora *et al.*	2019	Punjabi	Proceedings of Advances in Signal Processing and Communication	Punjabi dialect's Majhi Doabi Malwai	-	--	-	Mean pitch 370.23 162.23 194.190

Jaspreet Kaur
29/8/2019.

Scanned by CamScanner

3 Research Gaps

Lots of work has been done on ASR system for tonal languages like Chinese, Thai, Vietnamese and Mandarin but Punjabi ASR system is still at its initial stage of research. In implementing Punjabi ASR system, there are some of the issues related to language as well as related to region specific challenges. The following are the issues related to Punjabi ASR.

3.1 Lack of Speech Corpus in Punjabi Language

The initial step in implementing ASR system for any language is the building of the speech corpus in which the language linguistic details are contained. Furthermore, the ASR system performance depends on the speech corpus accuracy. Hence the two mandatory issues that has to be taken into account while building the speech corpus are accuracy and availability. The above mentioned issues are the major reasons for deficiency for Punjabi ASR research as compared to other tonal languages like Chinese, Thai, Vietnamese and Mandarin. We still lack in standard and accurate speech corpus for Punjabi. Few of researchers are working for building Punjabi speech corpus. However, these speech corpus are not available easily for public use.

3.2 Variation of Punjabi Accent Depending from Region to Region

Punjabi is spoken all over the globe. There are many Punjabi accents and dialects. Each and every region has its own Punjabi accents and dialects. From region to region the accent of Punjabi changes. The development of ASR system for Punjabi is not as easy to be done as other tonal languages, owing to its various accents.

3.3Focus on only Isolated and Connected Speech Corpus

For Punjabi ASR, the research on continuous as well as on spontaneous speech is still emerging. The major work has been done on isolated and connected speech corpus. Absence of effective acoustic and language modeling in Punjabi is noticed.

3.4 Unavailability of Commercial ASR system in Punjabi

Till date, no standard commercial ASR system is available for Punjabi. Even Google speak, Siri and other commercial software's do not support the language.

3.5 Tone Issues

a. In Indo-European language family, Punjabi shows tonal characteristics which are very different from other tonal languages.

Jaspreet Kaur
29/1/2019

Scanned by CamScanner

b. Punjabi tonal nature differs from other tonal languages like Chinese, Thai and Vietnamese as the tone of Punjabi depends on word as well as on character. Punjabi has five aspirated sounds which act as tonal characters and their pronunciation varies as per their position in the pronounced word and accordingly three phonemically different tones: high, low and level are produced. Punjabi has glottal sound /h/, which is used to generate tones when it is placed in the middle and end position of the word.

c. All dialects do not have the same tone as Lahnda does not have the tone whereas Eastern Punjabi dialects have the tones.

4 Problem Formulation

Punjabi has earned the rank of 10th most spoken language of the world but still is under resourced language as there is no standard speech corpus available for the language. Due to the tonal nature of Punjabi building ASR system is not simple. Till date, accurate tone is not embedded into system, due to that, high accuracy is not achieved. State- of-the-art for Chinese, Thai, English and Vietnamese is available. But state- of-the- art for Punjabi is not available being the 10th most spoken language of the world. The main purpose of this research is to design Punjabi ASR system embedded with tone characteristics so that it can work with more accuracy.

5 Research Objectives

The major objectives of this research work include the following:-

1. To collect and annotate speech corpus of Punjabi language having tonal characteristics.
2. To investigate and select an efficient feature extraction and classification technique(s) for speech recognition in Punjabi.
3. To propose a technique that optimizes tonal feature selection for Punjabi speech recognition.
4. To propose an efficient classifier to enhance the performance of the system.
5. To analyze and compare the performance of the proposed system.

6 Contribution of the Work to be Proposed

ASR system can also be utilized for operating mobile devices with hands-free environment. The research work will show the path that can lead to the ASR system for Punjabi and can help Punjabi people to use ASR system for many types of applications. Punjabi ASR system can provide help to farmers by taking input as a speech signals and converting them into text format for searching the output from the database and the pre-

Satpreet Singh
26/8/2019

Scanned by CamScanner

recorded outputs can be played back according to the question/query of the farmer. Punjabi ASR system can also be helpful for health, banking, education etc.

7 Methodology

The research methodology that will be focused in our research is shown in figure 1 and is explained step by step.

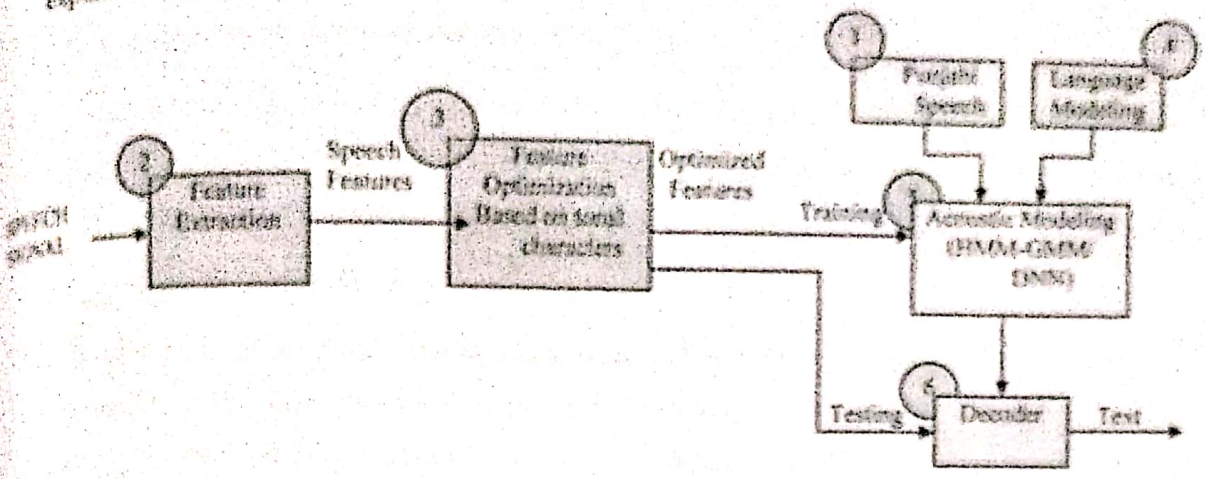


Figure 1: Flow Diagram of Proposed System

Step 1: Design of Corpus

A corpus of least of 50 hours containing the Punjabi high tone, low tone and level tone will be collected with the help of good quality microphone and recording device.

Step 2: Feature Extraction

In feature extraction step, processing of input speech signal is done to produce feature vectors which are non-redundant as well as unique in nature. The extracted features have the capability to distinguish each one from same kind of feature vectors. Knowledge regarding phonetic class is provided by processed vectors and the processed vectors remove the information part that does not have linguistic knowledge. Speakers or environment properties are also discarded by the processed vectors. In this phase, the tone features will be extracted using the existing feature extraction techniques, like MFCC, PLP, GFCC, Relative Spectral Transform (RASTA) and Linear Predictive Coefficient (LPC).

Step 3: Optimization Based on Tonal Characters

Various combinations of feature extraction techniques will be performed along with tone attributes to propose efficient feature extraction and selection techniques.

Scanned by CamScanner

Step 4: Design of Language Model

LM (Language Model) helps to improve the performance of ASR system. LM assists in selecting highest hypothesis of same type, typically a lattice or N-best list as recognition system's output. So, in our research statistical approach, for example n-gram LM, will be used to design LM.

Step 5: Designing Acoustic Model and its Training

Many statistical approaches are available for building acoustic model. HMM, DNN and GMM have been employed earlier. In our research, combination like HMM/GMM/DNN along with GA will be used to improve the performance.

Step 6: Comparison of Result

Firstly the baseline model will be designed and its performance will be noted for various parameters. After that, the same model will be trained using proposed techniques. The proposed system will be compared with baseline system and improvements will be notified for different parameters.

8 Year-Wise Plan of the Work & Targets to be Achieved

Table2: Time-Line Chart

Time Scale	1st Year				2nd Year				3rd Year			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Literature Survey												
Data collection												
Pre-processing												
Feature Extraction												
Feature Selection												
Classification/Recognition												

9 Place of Work

Maharaja Ranjit Singh Punjab Technical University, Bathinda.

10 Resources (Hardware/Software) Required & Availability Status

Linux Environment like Ubuntu 10.04.4 32 bit.
Audacity tool for recording of speech files 1.3.12 –beta/PRAAT.
Kaldi tool for the purpose of speech recognition.
MATLAB.

Jaspreet Kaur
29/8/2019

Scanned by CamScanner

11 Extent of External Assistance Required (if any)

No

Abbreviations

AMDF	Average Magnitude Difference Function
ASR	Automatic Speech Recognition
BLSTM	Bidirectional Long Short Term Memory
BNF	Bottle Neck Features
BTEC	Basic Travel Expression Corpus
CD	Context Dependent
CE	Consonantal Effects
CER	Character Error Rate
CI	Context Independent
CNN	Convolutional Neural Network
CRF	Conditional Random Field
CTC	Connectionist Temporal Classification
DBNFs	Deep Bottle Neck Features
DE	Differential Evolution
DNN	Deep Neural Network
DR	Detection Rate
EMG	Electromyography
ENV	Environment
GA	Genetic Algorithm
GFCC	Gammatone Frequency Cepstral Coefficient
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
LDA	Linear Discriminant Analysis
LM	Language Model
LPC	Linear Predictive Coefficient
LSTM	Long Short- Term Memory
LSTM-OP	Long Short- Term Memory with Output Projection Layer
LSTMP	Long Short- Term Memory Projected
LVCSR	Large Vocabulary Continuous Speech Recognition
MFCC	Mel Frequency Cepstral Coefficient
MFPLP	Mel Frequency with Perceptual Linear Prediction

Jaspreet Kaur
29/8/2019.

Scanned by CamScanner

MLE	Maximum Likelihood Estimation
MLLT	Maximum Likelihood Linear Transformation
MMI	Maximum Mutual Information
MSD	Multispace Distribution
NCC	Normalized Cross-Correlation
NN	Neural Network
PER	Phone Error Rate
PLP	Perceptual Linear Prediction
PNCC	Power Normalized Cepstral Coefficients
RASTA	Relative Spectral Transform
RNN	Recurrent Neural Network
SER	Sentence Error Rate
SGMM	Subspace Gaussian Mixture Model
SVM	Support Vector Machine
VOV	Voice of Vietnam
WA	Word Accuracy
WER	Word Error Rate

Jaspreet Kaur
29/8/2019

Scanned by CamScanner

References

Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C, Casper J, Catanzaro B, Cheng Q, Chen G, Chen J, Chen J, Chen Z, Chrazanowski M, Coates A, Diamos G, Ding Ke, DU N, Elsen E, Engel J, Fang W, Jiang B, Ju C ,Jun B, Legresley P, Lin L, Liu J, Liu Y, Li W, Li X, Ma D, Narang S, Ng A, Ozair S, Peng Y, Prenger R, Qian S, Srinet K, Sriram A, Tang H, Tang L, Wang C, Wang J, Wang K, Wang Yi, Wang Z, Wang Z, Wu S, Wei L, Xiao B, Xie W, Xie Y, Yogatama D, Yuan B,Zhan J, Zhu Z,2016 Deep Speech 2:End-to-End Speech Recognition in English and Mandarin. *Proceedings of the 33rd International Conference on Machine Learning (ICML), New York*, 48:173-182.

Arora A, Kadyan V, Singh A, 2019 Effect of Tonal Features on Various Dialectal Variations of Punjabi language. *Proceedings of the Conference on Advances in Signal Processing and Communication*, 467-475

Chan W, L I, 2016 On Online Attention-based Speech Recognition and Joint Mandarin Character-Pinyin Training. *Proceedings of the Interspeech, San Francisco USA*, 3404-3408.

Chen NF, Wee D, Tong R, Ma B,Li H, 2016 Large-Scale Characterization of Non-Native Mandarin Chinese Spoken by Speakers of European Origin: Analysis on iCALL. *Speech Communication*, 84:46-56.

Dey A, Lalhminghlui W, Sarmah P, Samudravijaya K, Prasanna SRM, Sinha R, Nirmala SR, 2017 Mizo Phone Recognition System. *Proceedings of the India Council International Conference (INDICON)*, 1-5.

Dua M, Aggarwal RK, Kadyan V, Dua S, 2012 Punjabi Speech to Text System for Connected Words. *Proceedings of the 4th International Conference on Advances in Recent Technologies in Communication and Computing*, 206-209.

Govind D, Sarmah P, Prasanna SRM, 2012 Role of Pitch Slope and Duration in Synthesized Mizo Tones. *Proceedings of the Speech Prosody* 2012

Guglani J, Mishra AN, 2018 Continuous Punjabi Speech Recognition Model Based on Kaldi ASR. *International Journal of Speech Technology*, 21(2):211-216.

Hu X, Saiko M, Hori C, 2014 Incorporating Tone Features to Convolutional Neural Network to Improve Mandarin/Thai Speech Recognition. *Proceedings of the Signal and Information Processing Association Annual Summit and Conference (APSIPA), Asia-Pacific,Siem Reap,Combodia,* 1-5 .

Huang H, Hu Y, Xu H, 2017 Mandarin Tone Modeling using Recurrent Neural Networks. *arXiv preprint arXiv:1711.01946.*

Huy NV, Mai LC, Thang VT, Truong DQ, 2014 Vietnamese Recognition using Tonal Phoneme Based on Multi Space Distribution. *Journal of Computer Science and Cybernetics*, 30(1):28-38.

Kadyan V, Mantri A, Aggarwal RK, 2017a A Heterogeneous Speech Feature Vectors Generation Approach with Hybrid HMM Classifiers. *International Journal of Speech Technology*, 20(4): 761-769.

Kadyan V, Mantri A, Aggarwal RK, 2017b Refinement of HMM Model Parameters for Punjabi Automatic Speech Recognition (PASR) System. *IETE Journal of Research*, 64(5):673-688.

Kadyan V, Mantri A, Aggarwal RK, Singh A, 2018 A Comparative Study of Deep Neural Network Based Punjabi –ASR System. *International Journal of Speech Technology*, 22(1):111-119.

Kaur A, Singh A, 2016a Power-Normalized Cepstral Coefficients (PNCC) for Punjabi Automatic Speech Recognition Using Phone Based Modelling in HTK. *Proceedings of the 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, India,* 372-375.

Kaur A, Singh A, 2016b Optimizing Feature Extraction Techniques Constituting Phone Based Modelling on Connected Words for Punjabi Automatic Speech Recognition. *Proceedings of the*

Jaspreet Kaur 29/8/2019

Scanned by CamScanner

2nd International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, 2104-2108.

Kertkeidkachorn N, Chanlekha H, Suri T, Likitsupin K, Vorapatratorn S, Hirankan P, Limpanadusadee W,Chuetanapinyo S, Plitakpawatkul K, Puangsri N, Tangsirirat N, Trakulsuk K, Punyabukkana P, Suchato A, 2012 The CU-MFEC Corpus for Thai and English Spelling Speech Recognition. *Proceedings of the International Conference on Speech Database and Assessments, Macau, China*, 18-23.

Lata S, Arora S, 2012 Exploratory Analysis of Punjabi Tones in Relation to Orthographic Characters: A Case Study. *Proceedings of the Workshop on Indian Language and Data: Resources and Evaluation Workshop Programme*, 76-80.

Lata S, Arora S, 2013 Laryngeal Tonal Characteristics of Punjabi –An Experimental Study. *Proceedings of the International Conference on Human Computer Interactions (ICHCI), Chennai, India*, 1-6.

Li J, Zhang H, Cai X, Xu Bo, 2015a Towards End-to End Speech Recognition for Chinese Mandarin using Long Short-Term Memory Recurrent Neural Networks. *Proceedings of the 16th Annual Conference of International Speech Communication Association, Dresden, Germany*, 3615-3619.

Li X, Wu X, 2015 Constructing Long Short-Term Memory Based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, Australia*, 4520-4524.

Li X, Yang Y, Pang Z, Wu X, 2015b A Comparative Study on Selecting Acoustic Modeling Units in Deep Neural Networks based Large Vocabulary Chinese Speech Recognition. *Neurocomputing*, 170:251-256.

Nguyen QB, Vu TT, Luong CM, 2016 The Effect of Tone Modeling in Vietnamese LVCSR System. *Procedia Computer Science*, 81:174-181.

Nguyen TL, Tran DD, 2012 Influences of Particles on Vietnamese Tonal Co-articulation. *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), Mumbai*, 163-172.

Nguyen VH, Luong CM, Vu TT, 2015 Tonal Phoneme Based Model for Vietnamese LVCSR. *Proceedings of the International Conference Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), Shanghai, China*, 118-122.

Sarma BD ,Sarmah P , Lalhminghlui W, Prasanna SRM, 2015 Detection of Mizo Tones. *Proceedings of the 16th Annual Conference of the International Speech Communication Association, Dreseden, Germany*, 934 -937.

Shan C, Zhang J, Wang Y, Xie L, 2018 Attention-Based End-to-End Speech Recognition on Voice Search. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calagary, Canada*, 4764-4768.

Srijiranon K, Eiamkanitchat N, 2015 Thai Speech Recognition Using Neuro-Fuzzy System. *Proceedings of the 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Hua Hin, Thailand*,1-6.

SrisuwanN, Phukpattaranont P, Limsakul C, 2013 Three Steps of Neuron Network Classification for EMG-Based Thai Tones Speech Recognition. *Proceedings of the 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology,Krabi, Thailand, 1-6.*

Theera-Umpon N, Chansareewittaya S, Auephanwiriyakul S, 2011 Phoneme and Tonal Accent Recognition for Thai Speech. *Expert Systems with Applications*, 38(10):13254-13259.

Yang D, Pan YC, Furui S, 2012 Vocabulary Expansion Through Automatic Abbreviation Generation for Chinese Voice search. *Computer Speech and Language*, 26(5):321-335.

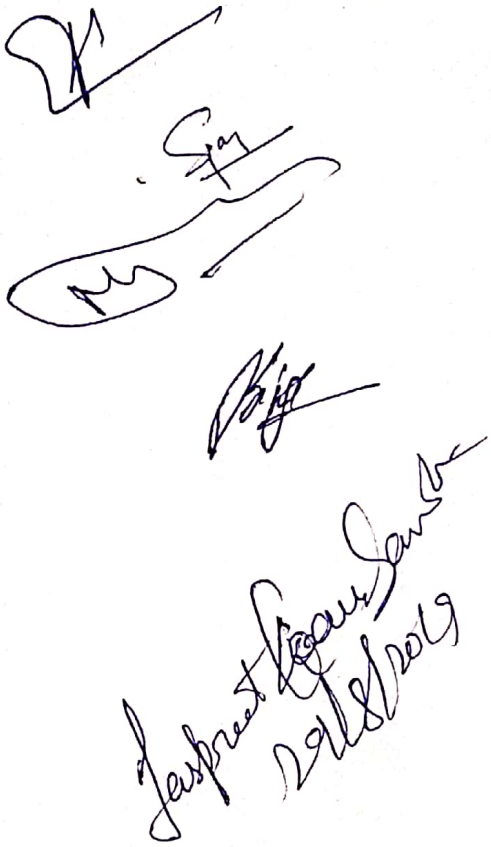
Jaspreet Kaur Sandhu 29/8/2019

Scanned by CamScanner

Zhou S, Dong L, Xu S, Xu B, 2018a Syllable-Based Sequence-to Sequence Speech Recognition with the Transformer in Mandarin Chinese. *arXiv preprint arXiv:1804.10752.*

Zhou S, Dong L, Xu S, Xu B, 2018b A Comparison of Modeling Units in Sequence-to- Sequence Speech Recognition with the Transformer on Mandarin Chinese. *Proceedings of the International Conference on Neural Information Processing (ICONIP), 210-220.*

Zou W, Jiang D, Zhao S, Li X, *2018* A Comparable Study of Modeling Units for End-to-End Mandarin Speech Recognition. *arXiv preprint arXiv:1805.03832.*



Scanned by CamScanner